Large Language Model Unlearning

Year: 2023 | Citations: 217 | Authors: Yuanshun Yao, Xiaojun Xu, Yang Liu

---

Abstract

We study how to perform unlearning, i.e. forgetting undesirable misbehaviors, on large language models (LLMs). We show at least three scenarios of aligning LLMs with human preferences can benefit from unlearning: (1) removing harmful responses, (2) erasing copyright-protected content as requested, and (3) reducing hallucinations. Unlearning, as an alignment technique, has three advantages. (1) It only requires negative (e.g. harmful) examples, which are much easier and cheaper to collect (e.g. via red teaming or user reporting) than positive (e.g. helpful and often human-written) examples required in RLHF (RL from human feedback). (2) It is computationally efficient. (3) It is especially effective when we know which training samples cause the misbehavior. To the best of our knowledge, our work is among the first to explore LLM unlearning. We are also among the first to formulate the settings, goals, and evaluations in LLM unlearning. We show that if practitioners only have limited resources, and therefore the priority is to stop generating undesirable outputs rather than to try to generate desirable outputs, unlearning is particularly appealing. Despite only having negative samples, our ablation study shows that unlearning can still achieve better alignment performance than RLHF with just 2% of its computational time.